Equivariant Graph Neural Networks for Toxicity Prediction

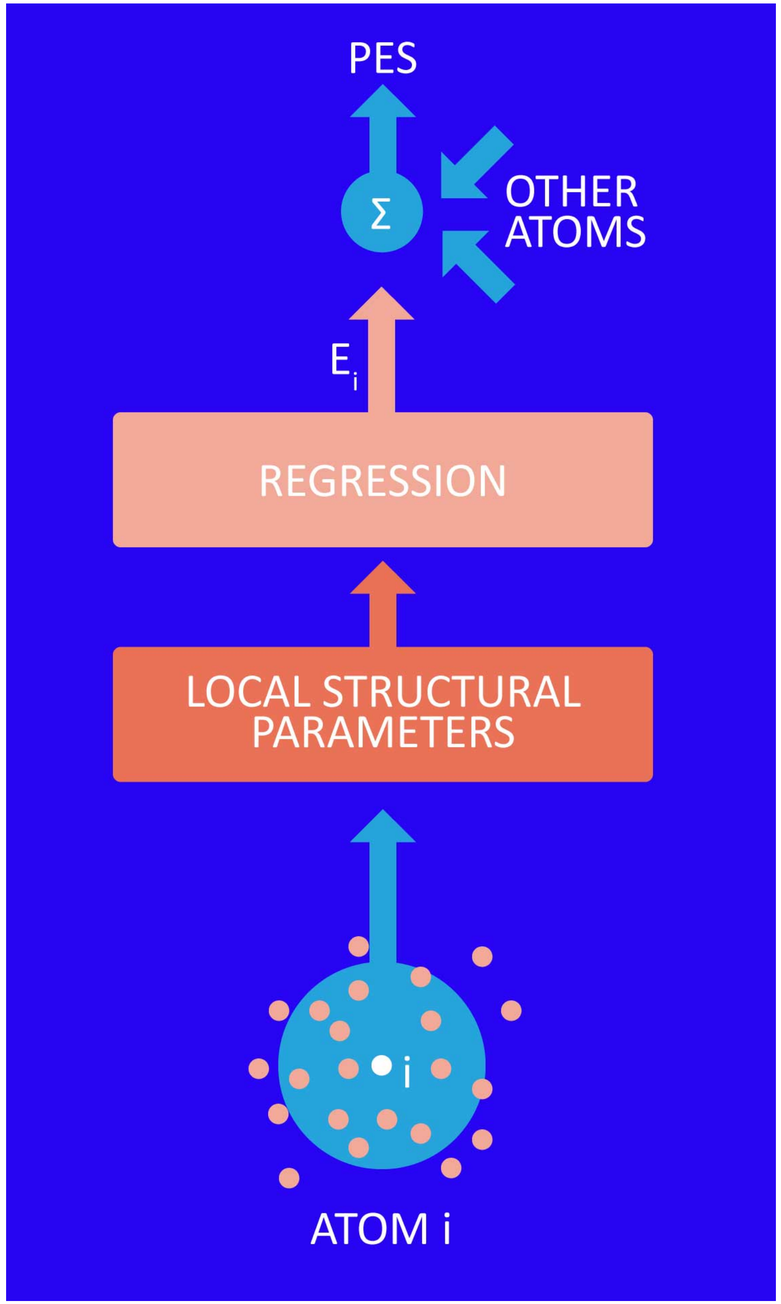
AIDD TALK| 02-2023

Universitat Pompeu Fabra Barcelona

INTRODUCTION

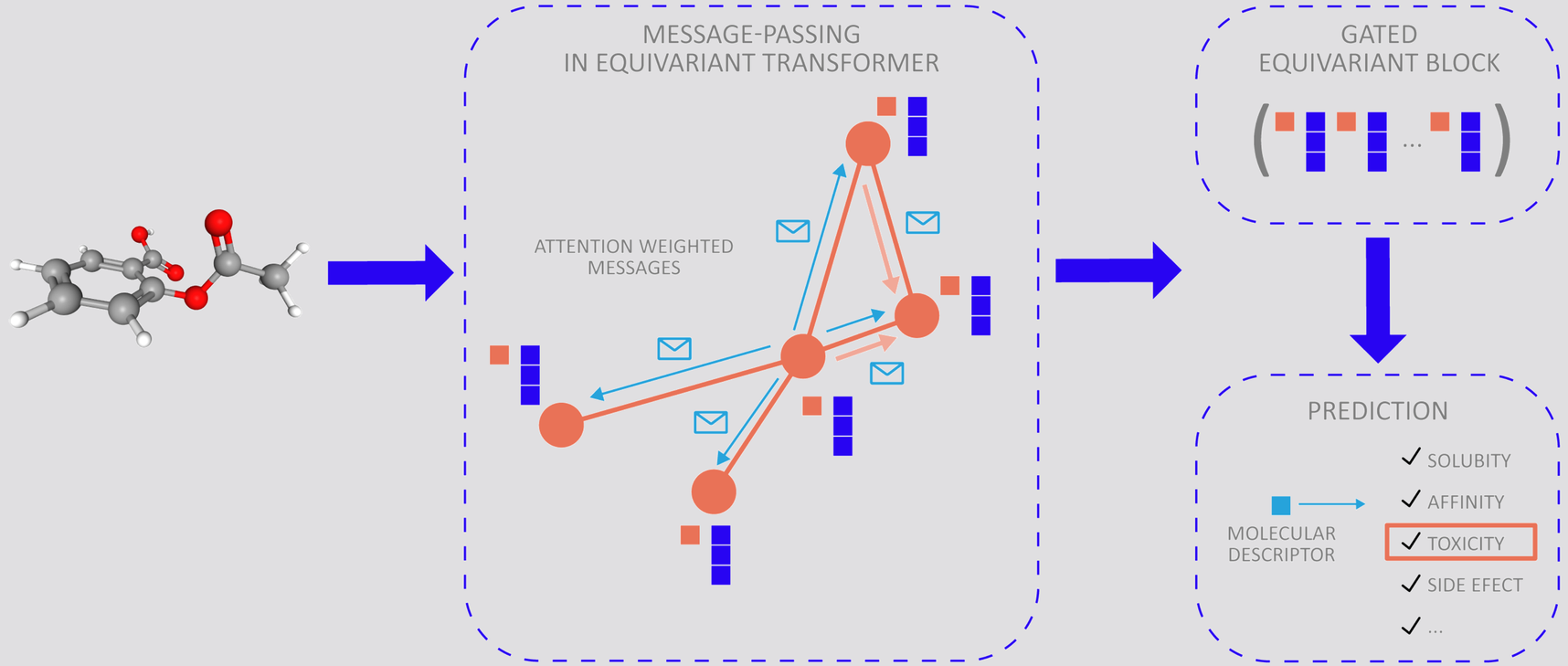
Research Overview

- Machine Learning Force Fields and Interatomic Potentials to study dynamics of molecules and materials
- Learning the PES from quantum mechanical reference data via Neural Network approximation
- But, molecules live in 3D (Euclidean/physical) space
  - MLFFs need to be able to “understand” symmetry/isometries of Euclidean space
    - Energy invariance/force equivariance
- Importantly, learn meaningful, equivariant atom representations based on 3D point cloud
  - **Use equivariant graph neural networks**

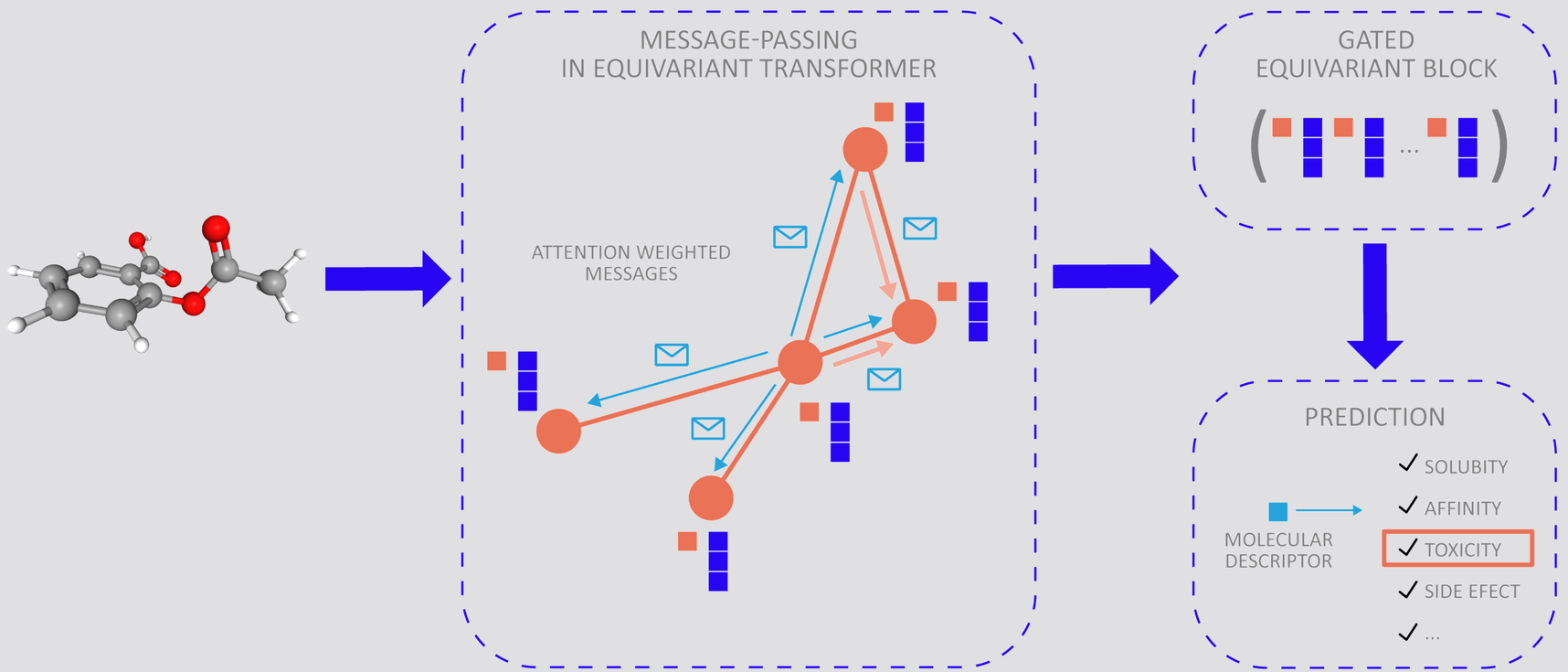


Toxicity Prediction - Overview

- We use 3D geometries of molecules to encode them via an equivariant graph (message-passing) neural network
  - Atomic representations are then used to predict pharmacokinetic or toxicological profile of the molecule

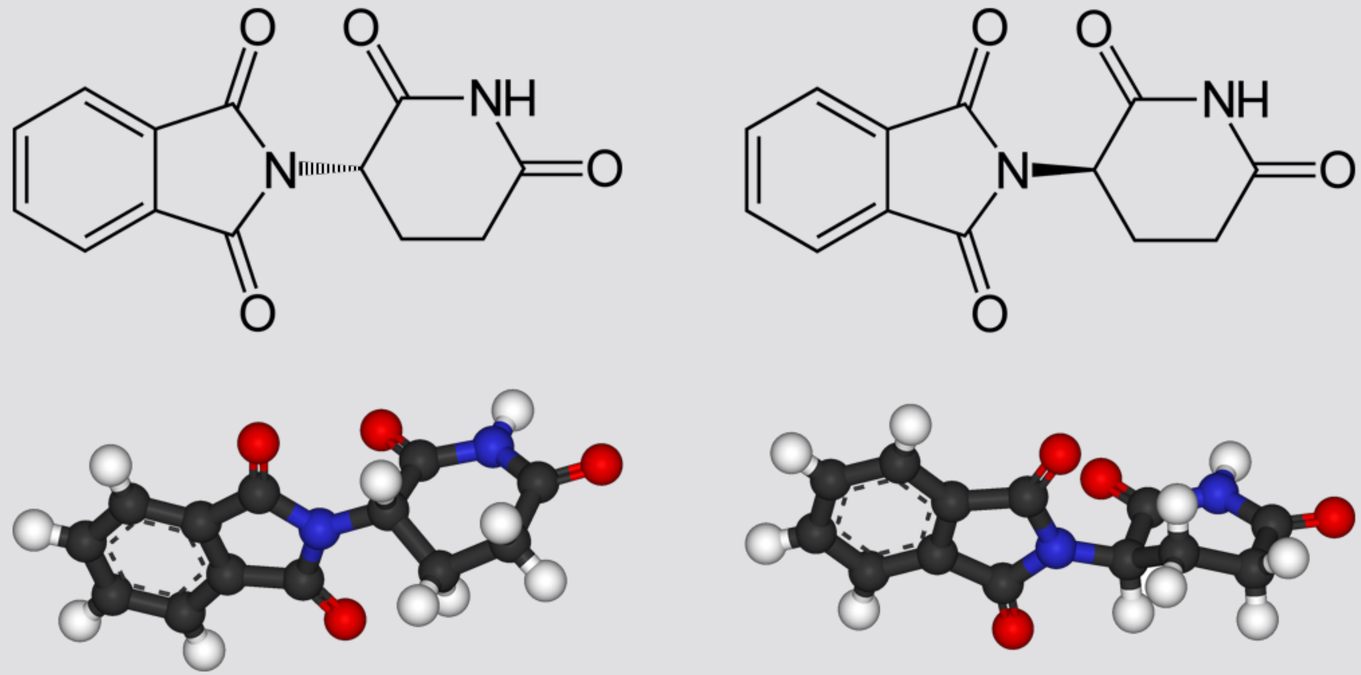


Toxicity Prediction - Overview



Toxicity Prediction - why 3D?

- Important ADMET profiles might not be correctly covered by other representations like SMILES strings or fingerprints (the larger the system)
  - Examples:
    - Cis-Platin / Trans-Platin
    - Thalidomide



S-enantiomer (embryo-toxic)

R-enantiomer (sedative)

METHODS

Methods

MESSAGE-PASSING NEURAL NETWORKS: OVERVIEW

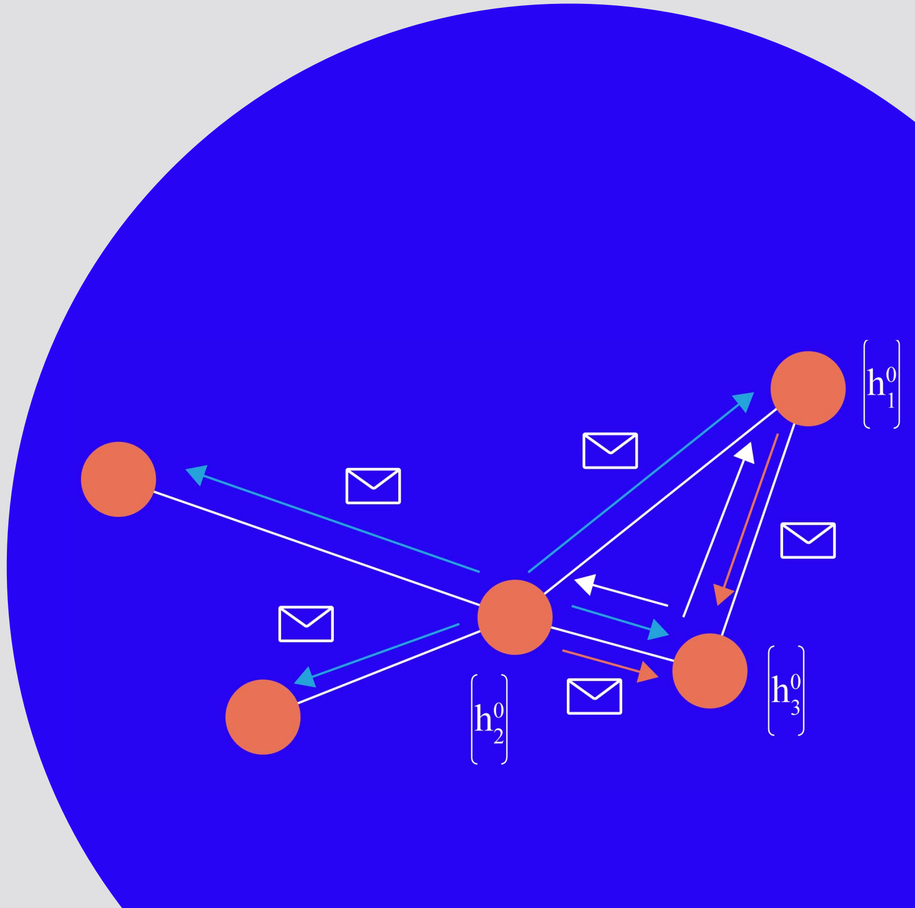
- We construct molecules as graphs
  - ➡ Each node i has a hidden state $h_i^{t,l}$ at iteration t with order l

- We collect messages from neighboring nodes/atoms j in a given cutoff region with an edge index e_{ij}:

$$m_i^{t+1} = \sum_{j \in N(i)} M_t(h_i^t, h_j^t, e_{ij})$$

- We can additionally add attention weights to every message, such that the model can learn to prioritize

- We update every node: $h_i^{t+1} = S_t(h_i^t, m_i^{t+1})$



Methods

SYMMETRY: EQUIVARIANCE

- [We deal only with isometries of the Euclidean space]

- Given a set of transformations $T_g : \mathcal{V} \to \mathcal{V}$ for $g \in G$
- Given a function $f : \mathcal{V} \to \mathcal{W}$ (in our case NN)

- Equivariance of f:

$$\exists S_g : \mathcal{W} \to \mathcal{W} : S_g[f(\mathbf{x})] = f(T_g[\mathbf{x}]) \ \forall g \in G, \ \mathbf{x} \in \mathcal{V}$$

  - ➡ Example: Spherical harmonics
    $\mathbf{Y}_J : S^2 \to \mathbb{C}^{2J+1}$ for $J \geq 0$ are equivariant to SO(3) (group of 3D rotations)
  - ➡ $\mathbf{Y}_J(\mathbf{R}_g^{-1}\mathbf{x}) = \mathbf{D}_J^*(g)\mathbf{Y}_J(\mathbf{x})$, where $\mathbf{x} \in S^2, \ g \in G$

Methods

EQUIVARIANCE: BENEFITS

- Molecules come with invariance of energy towards permutation of atom indices, global rotation and translation
  - ➡Equivariance of tensorial properties, like forces, multipole expansion of electron density etc.

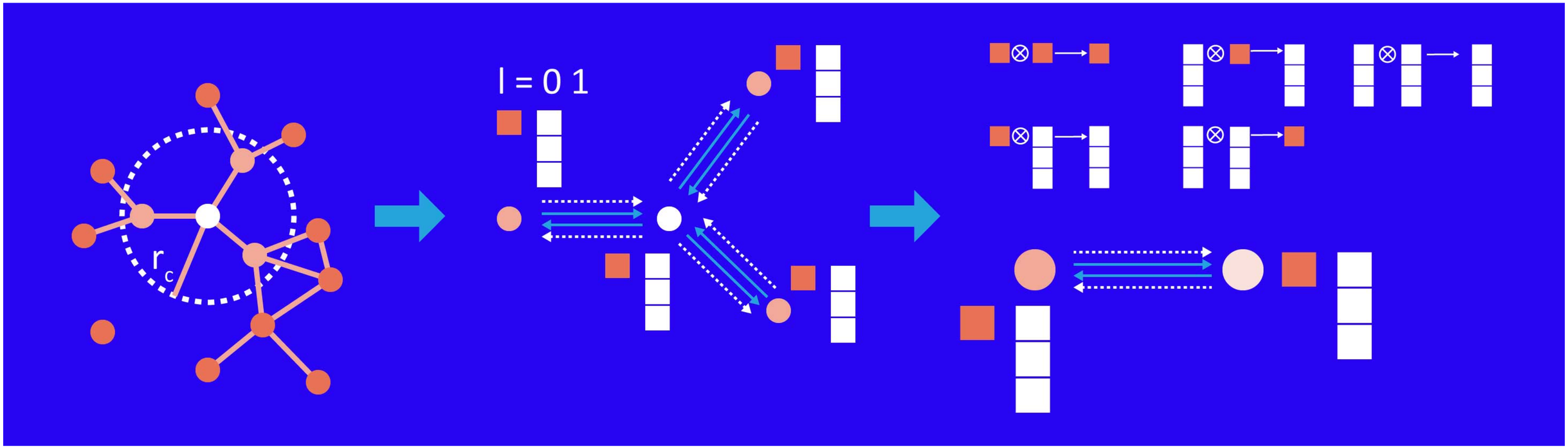
- No data augmentation: data efficiency
- Restricting functional space: learning efficiency

- Introduce directional, equivariant information (inductive bias)
  - ➡ Rotation-invariant representations (scalar reps.) do not propagate well directional information

- Possibility to predict tensorial properties, not only scalar output

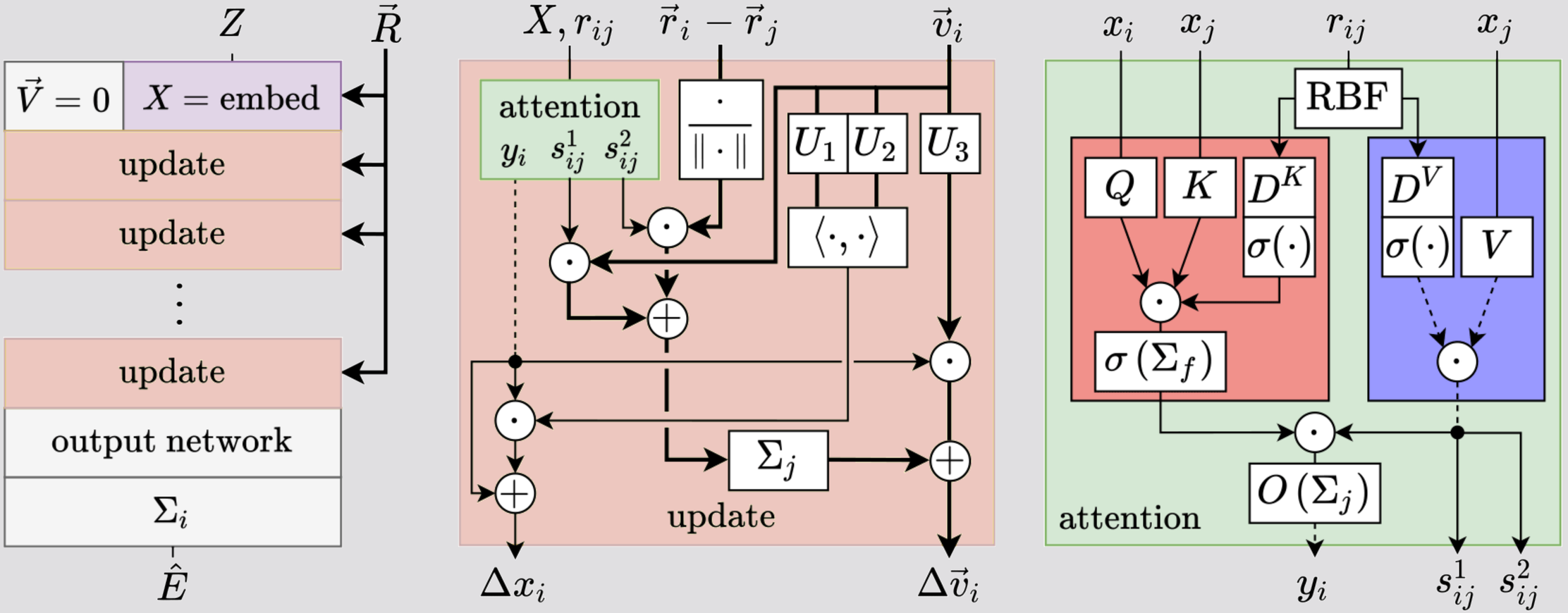
Methods

EQUIVARIANT MESSAGE-PASSING NEURAL NETWORKS: OVERVIEW



Methods

TORCHMD-NET: EQUIVARIANT TRANSFORMER



Thoelke and Fabritiis, arXiv:2202.02541 (2021)

RESULTS

Toxicity Datasets

- MoleculeNet: Tox21, ToxCast, SIDER, ClinTox, BACE, BBBP
- TDCommons: Ames, hERG, DILI, Skin Reaction, LD50
- ToxBenchmark: Ames mutagenicity

  - 3D conformers for MoleculeNet taken from GEOM dataset
  - 3D conformer generation using CREST and GFN2-xTB with an implicit solvation model for TDCommons and ToxBenchmark

Toxicity Datasets

Dataset	Property	Tasks	Species	Recovered
Tox21	Qualitative toxicity	12	7,677	98.0%
ToxCast	Qualitative toxicity	617	8,405	98.0%
SIDER	Drug side effects	27	1,356	95.1%
ClinTox	Toxicity of failed approved drugs	2	1,438	98.7%
BACE	BACE-1 inhibition	1	1,511	99.9%
BBBP	Blood-brain barrier penetration	1	1,959	99.2%
Ames	Mutagenicity	1	7,269	99.8%
hERG	Coordination of heart beating	1	650	99.2%
DILI	Drug-induced liver injury	1	470	98.9%
Skin Reaction	Skin sensitization	1	403	99.7%
LD50	Acute toxicity	1	7,353	99.5%
ToxBenchmark	Mutagenicity	1	6,489	99.6%

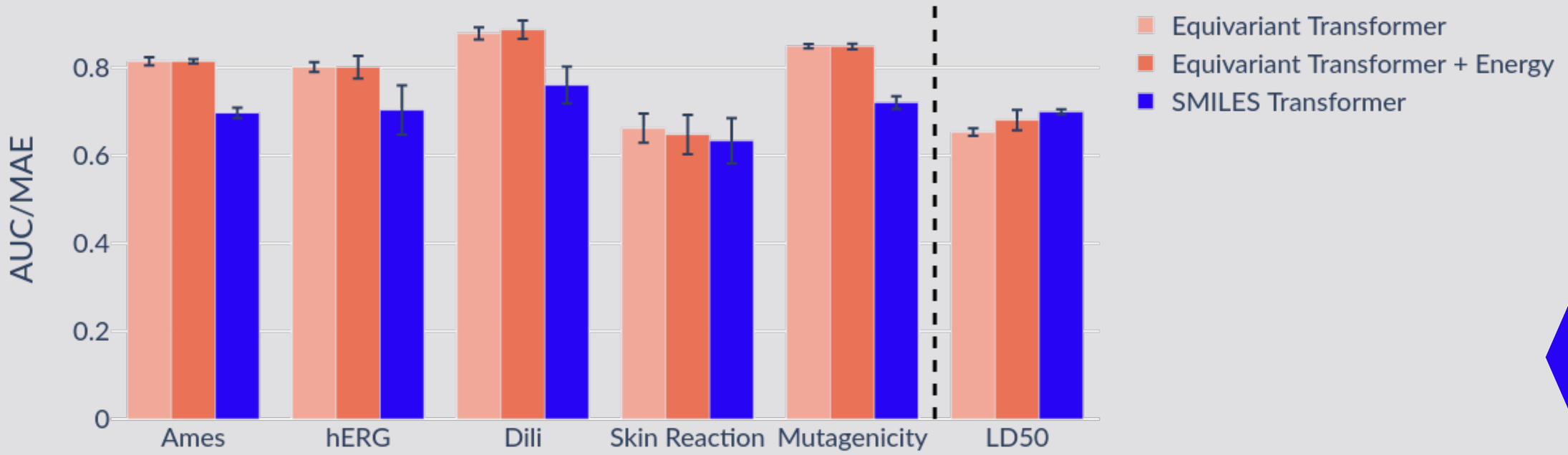
Benchmark: TDC and ToxBenchmark

OVERVIEW AND COMPARISON

Dataset	Ames ↑	hERG ↑	DILI ↑	Skin Reaction ↑	LD50 ↓	Tox-Benchmark ↑
No. molecules	7,269	650	470	403	7,353	6,489
AttrMasking	$0.842_{\pm 0.008}$	$0.778_{\pm 0.046}$	$0.919_{\pm 0.008}$	-	$0.685_{\pm 0.025}$	-
AttentiveFP	$0.814_{\pm 0.008}$	$0.825_{\pm 0.007}$	$0.886_{\pm 0.015}$	-	$0.678_{\pm 0.012}$	-
Fingerprint-based	$\mathbf{0.865}_{\pm 0.002}$	$\mathbf{0.875}_{\pm 0.003}$	$\mathbf{0.937}_{\pm 0.004}$	-	$\mathbf{0.588}_{\pm 0.005}$	$0.86_{\pm 0.01}$
SMILES-T	$0.697_{\pm 0.011}$	$0.703_{\pm 0.056}$	$0.760_{\pm 0.041}$	$0.633_{\pm 0.051}$	$0.715_{\pm 0.012}$	$0.720_{\pm 0.014}$
ET (single)	$0.836_{\pm 0.003}$	$0.839_{\pm 0.017}$	$0.878_{\pm 0.013}$	$0.662_{\pm 0.033}$	$0.653_{\pm 0.008}$	$\mathbf{0.881}_{\pm 0.008}$
ET (multi)	$0.804_{\pm 0.004}$	$0.763_{\pm 0.021}$	$0.885_{\pm 0.030}$	$0.581_{\pm 0.055}$	$0.660_{\pm 0.01}$	$\mathbf{0.879}_{\pm 0.005}$

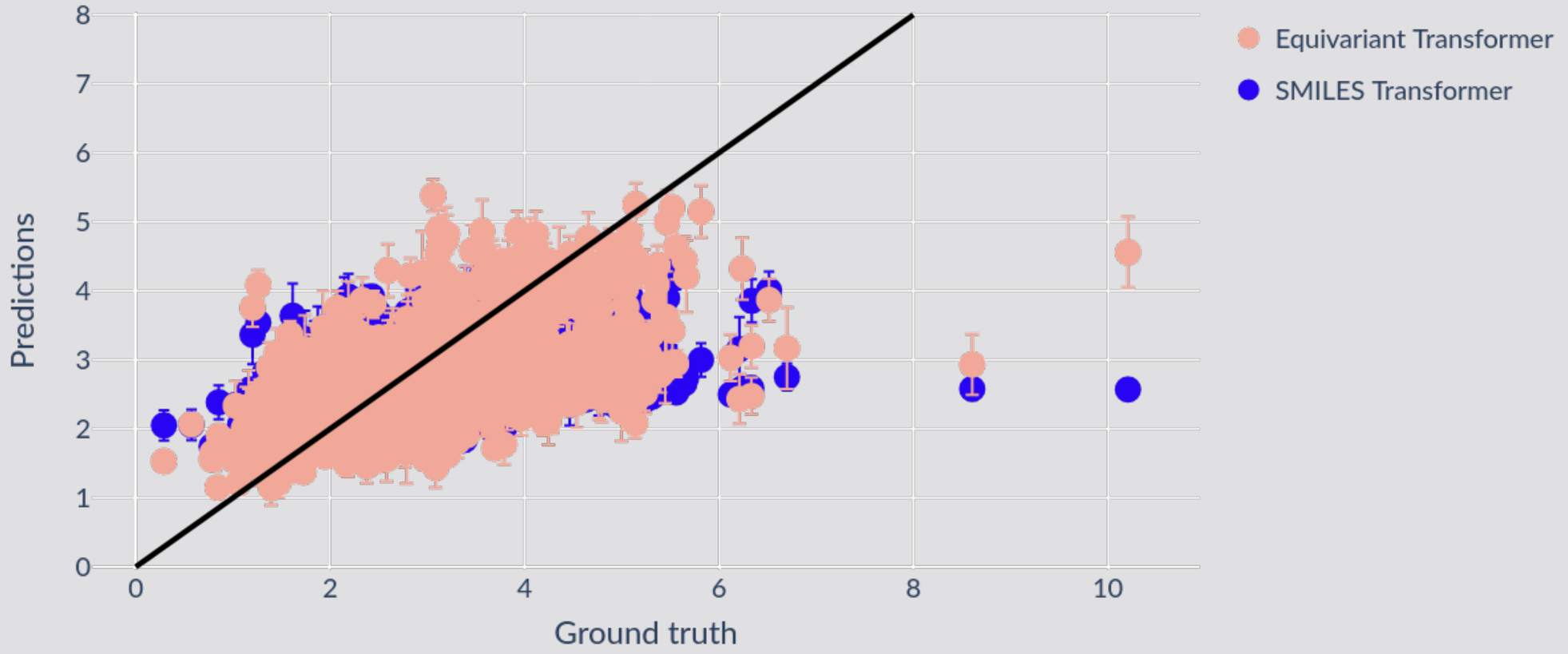
Benchmark: TDC and ToxBenchmark

ET VS. SMILES



Benchmark: TDC and ToxBenchmark

ET VS. SMILES



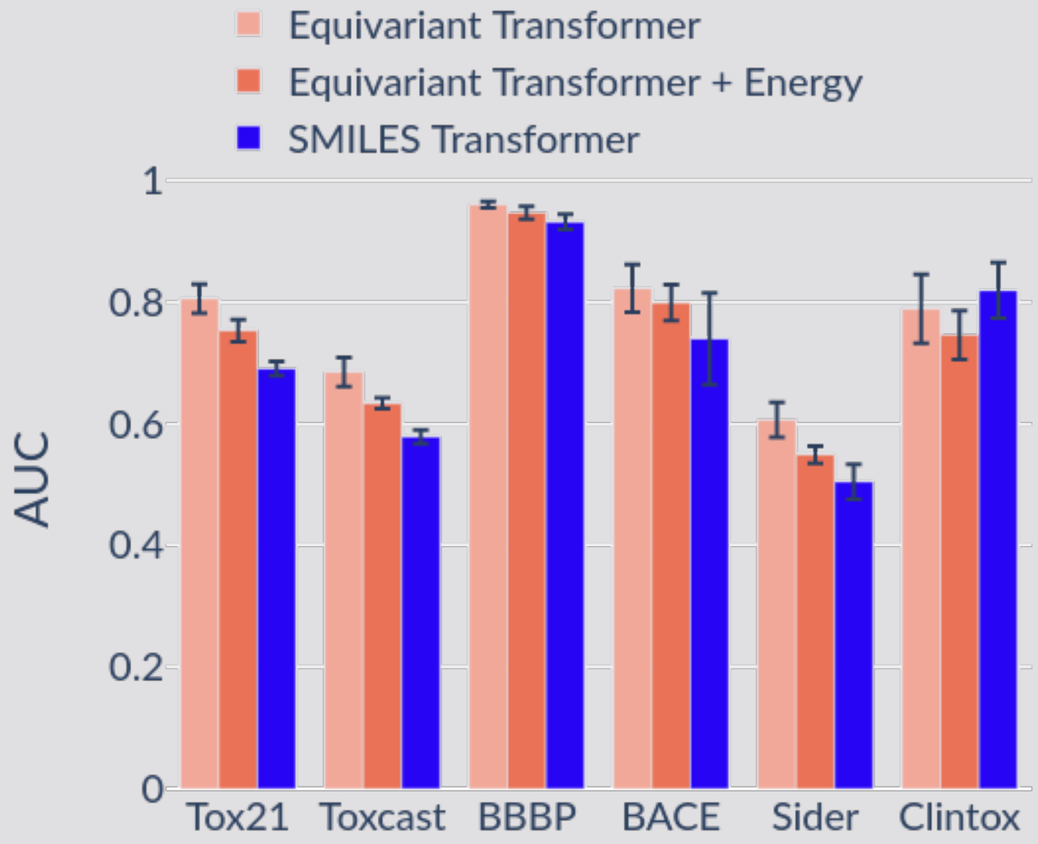
Benchmark: MoleculeNet

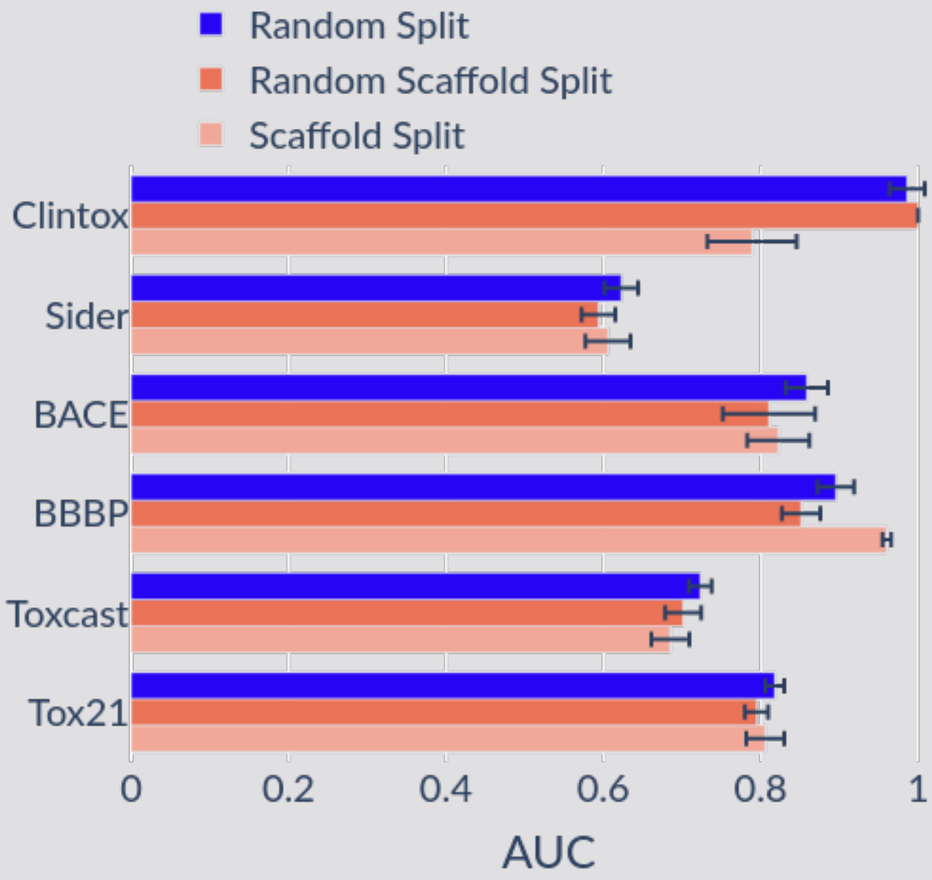
OVERVIEW AND COMPARISON

Dataset ↑	Tox21	ToxCast	SIDER	ClinTox	BACE	BBBP
No. molecules	7,677	8,405	1,356	1,438	1,511	1,959
No. tasks	12	617	27	2	1	1
D-MPNN	$0.759_{\pm 0.007}$	$0.655_{\pm 0.003}$	$0.57_{\pm 0.007}$	$\mathbf{0.906}_{\pm 0.006}$	$0.809_{\pm 0.006}$	$0.724_{\pm 0.004}$
AttentiveFP	$0.761_{\pm 0.005}$	$0.637_{\pm 0.002}$	$0.606_{\pm 0.032}$	$0.847_{\pm 0.003}$	$0.784_{\pm 0.022}$	$0.643_{\pm 0.018}$
GEM	$\mathbf{0.781}_{\pm 0.001}$	$\mathbf{0.692}_{\pm 0.004}$	$\mathbf{0.672}_{\pm 0.004}$	$0.901_{\pm 0.013}$	$\mathbf{0.856}_{\pm 0.011}$	$0.724_{\pm 0.004}$
SMILES-T	$0.691_{\pm 0.011}$	$0.578_{\pm 0.011}$	$0.504_{\pm 0.028}$	$0.819_{\pm 0.045}$	$0.739_{\pm 0.075}$	$0.931_{\pm 0.012}$
ET (single)	$\mathbf{0.805}_{\pm 0.024}$	$\mathbf{0.685}_{\pm 0.009}$	$0.606_{\pm 0.01}$	$0.851_{\pm 0.027}$	$0.832_{\pm 0.009}$	$\mathbf{0.960}_{\pm 0.03}$
ET (multi)	$0.751_{\pm 0.008}$	$0.623_{\pm 0.008}$	$0.560_{\pm 0.011}$	$0.843_{\pm 0.012}$	$0.816_{\pm 0.013}$	$\mathbf{0.955}_{\pm 0.008}$

Benchmark: MoleculeNet

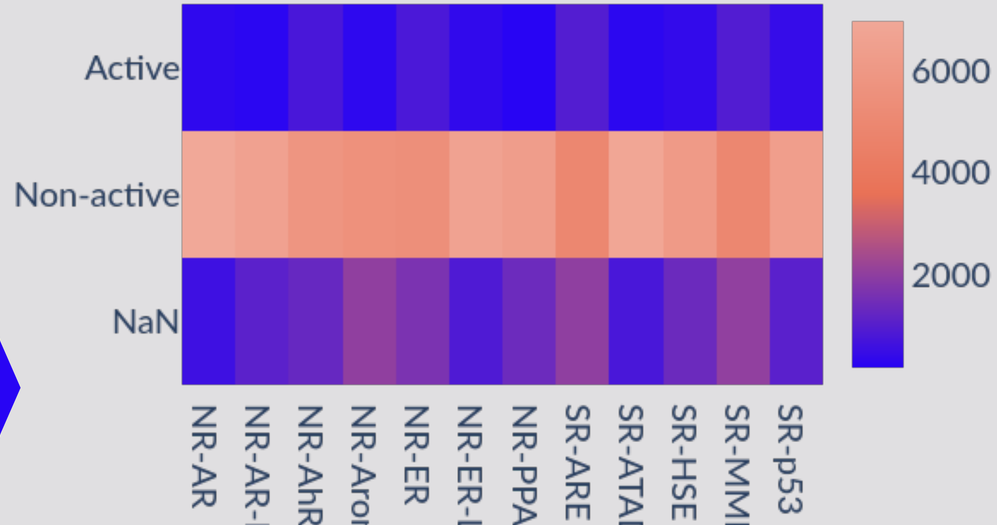
ET VS. SMILES

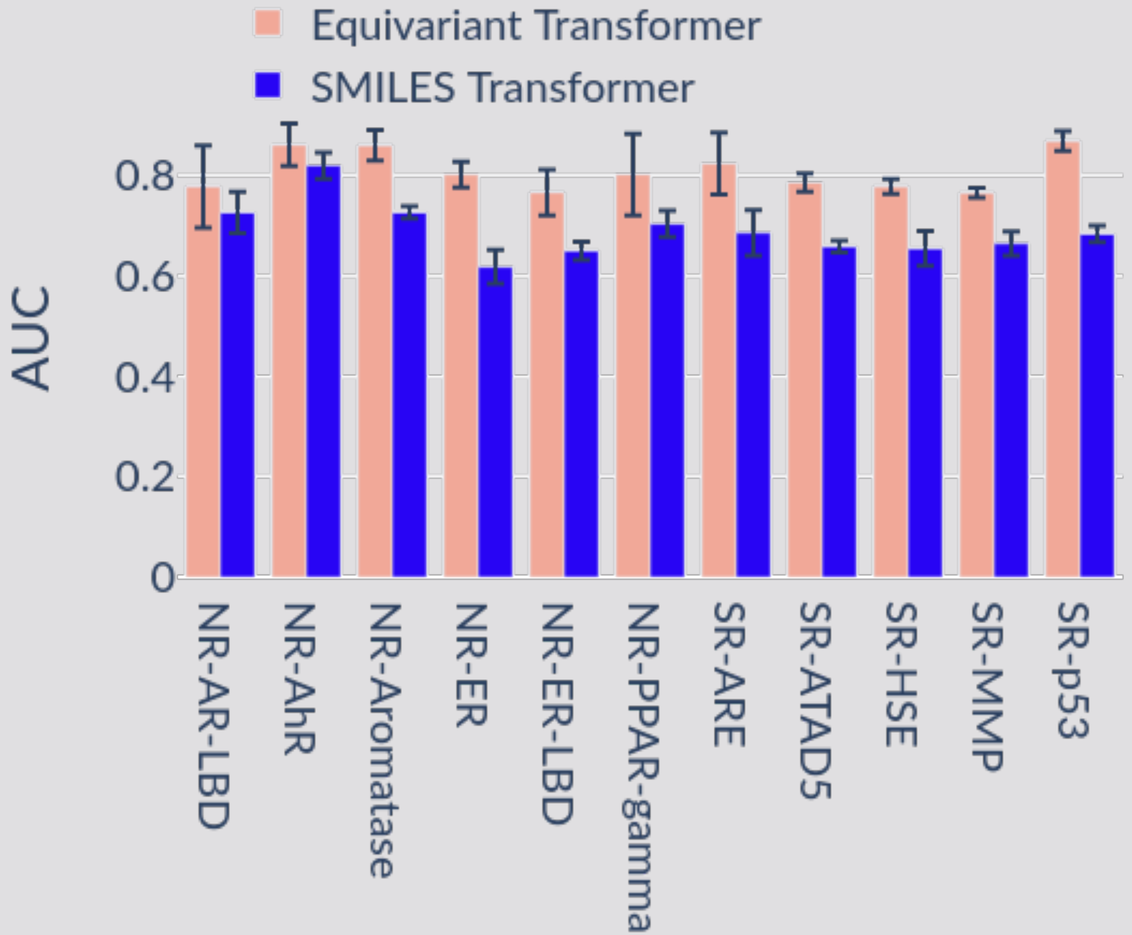




Benchmark: MoleculeNet

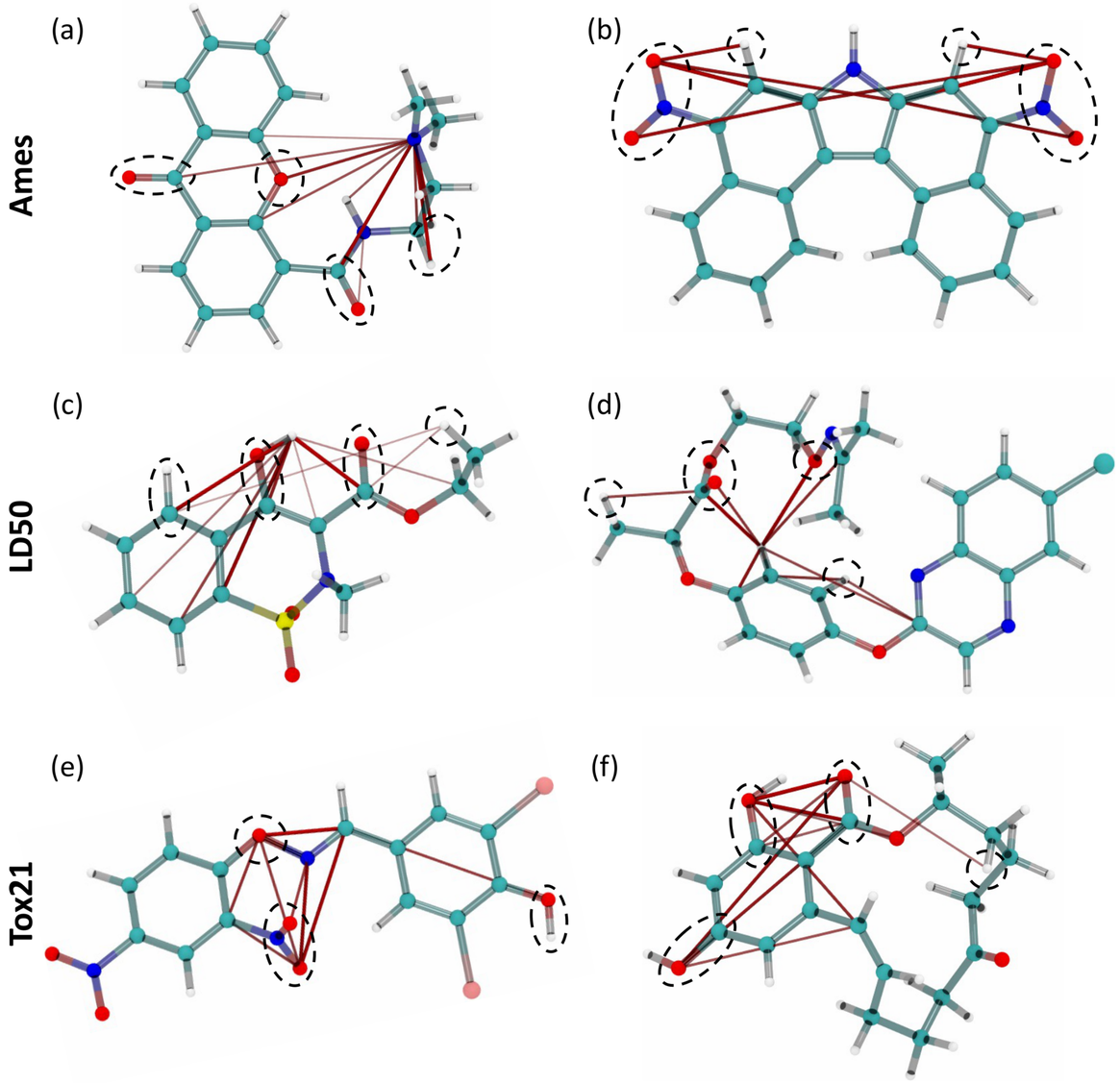
INVESTIGATION OF TOX21





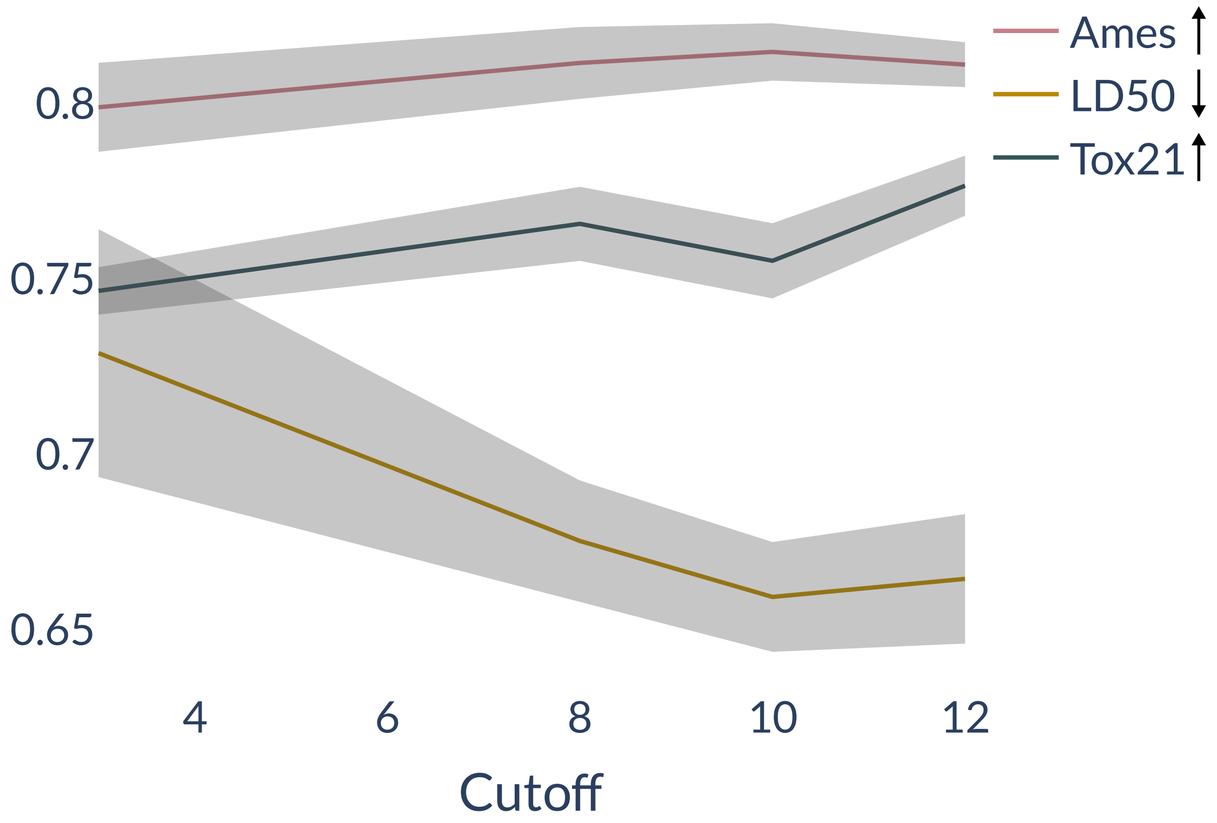
Attention weights analysis

INVESTIGATION OF AMES, LD50 AND TOX21



Short/long range analysis

INVESTIGATION OF AMES, LD50 AND TOX21



THANKS!